An Efficient Cross-lingual Model for Sentence Classification Using Convolutional Neural Network

Yandi Xia[1], **Zhongyu Wei**[12], **Yang Liu**[1]
[1]Computer Science Department, The University of Texas at Dallas
Richardson, Texas 75080, USA
[2]School of Data Science, Fudan University, Shanghai, P.R.China
{yandixia,zywei,yangl}@hlt.utdallas.edu

Abstract

In this paper, we propose a cross-lingual convolutional neural network (CNN) model that is based on word and phrase embeddings learned from unlabeled data in two languages and dependency grammar. Compared to traditional machine translation (MT) based methods for cross lingual sentence modeling, our model is much simpler and does not need parallel corpora or language specific features. We only use a bilingual dictionary and dependency parser. This makes our model particularly appealing for resource poor languages. We evaluate our model using English and Chinese data on several sentence classification tasks. We show that our model achieves a comparable and even better performance than the traditional MT-based method.

1 Introduction

With the rapid growth of global Internet, huge amounts of information are created in different languages. It is important to develop cross-lingual NLP systems in order to leverage information from other languages, especially languages with rich annotations. Traditionally, cross-lingual systems rely highly on machine translation (MT) systems (Wan et al., 2011; Wan, 2011; Rigutini et al., 2005; Ling et al., 2008; Amini et al., 2009; Guo and Xiao, 2012; Chen and Ji, 2009; Duh et al., 2011). They translate data in one language into the other, and then apply monolingual models. One problem of such cross-lingual systems is that there is hardly any decent MT system for resource-poor languages. Another problem is the lack of high quality parallel corpora for resource-poor languages, which is required by MT systems.

Other work tried to address these problems by developping language independent representation learning and structural correspondence learning (SCL) (Prettenhofer and Stein, 2010; Xiao and Guo, 2013). They showed some promising results on document level classification tasks. However, their methods require carefully designed language specific features and find the "pivot features" across languages, which can be very expensive and inefficient.

To solve these problems, we develop an efficient and feasible cross-lingual sentence model that is based on convolutional neural network (CNN). Sentence modeling using CNN has shown its great potential in recent years (Kalchbrenner et al., 2014; Kim, 2014; Ma et al., 2015). One of the advantages is that CNN requires much less expertise knowledge than traditional feature based models. The only input of the model, word embeddings, can be learned automatically from large unlabeled text data.

There are roughly two main differences between different languages, lexicon and grammar. Lexicon can be seen as a set of symbols with each symbol representing certain meanings. A bilingual dictionary easily enables us to map from one symbol set to another. As for grammar, it decides the organization of lexical symbols, i.e., word order. Different languages organize their words in different manners (see Figure 1a for an example). To reduce grammar difference, we propose to use dependency grammar as an intermediate grammar. As shown in Figure 1b, dependency grammar can yield a similar dependency tree between two sentences in different languages.

To bridge two different languages from aspects of both lexicon and grammar, our CNN-based cross-lingual model consists of two components, bilingual word embedding learning and CNN incorporating dependency information. We propose

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics – Student Research Workshop, pages 126–131,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

a method to learn bilingual word embeddings as the input of CNN, using only a bilingual dictionary and unlabeled corpus. We then adopt a dependency-based CNN (DCNN) (Ma et al., 2015) to incorporate dependency tree information. We also design lexical features and phrase-based bilingual embeddings to improve our cross-lingual sentence model.

We evaluate our model on English and Chinese data. We train a cross-lingual model on English data and then test it on Chinese data. Our experiments show that compared to the MT based cross-lingual model, our model achieves a comparable and even better performance on several sentence classification tasks including question classification, opinion analysis and sentence level event detection.

2 Methods

Our method is based on the CNN sentence classification model. It consists of two key components. First, we propose a method to learn bilingual word embeddings with only a bilingual dictionary and unlabeled corpus. This includes both word and phrase based embeddings. Second, for the CNN model, we use dependency grammar as the intermediate grammar, i.e., dependency-based CNN (DCNN) (Ma et al., 2015) where we also propose some useful modifications to make the model more suitable for the cross-lingual tasks.

2.1 Bilingual word and phrase embeddings

In order to train the bilingual word embeddings, we first construct an artificial bilingual corpus containing mix-language texts. We assume that the embeddings for a word and its translation in another language should be similar. We thus aim to create a synthetic similar context for a bilingual word pair. For example, assume we have an English unlabeled corpus and we want to learn word embeddings of Chinese word "夏威夷" and its English counter-part "Hawaii", we can substitute half of "Hawaii" in the English corpus into "夏威夷". Based on the modified corpus, we can obtain similar embeddings for the bilingual word pair "Hawaii" and "夏威夷". Similarly, we can also substitute Chinese words in the Chinese unlabeled data with their English counter-parts.

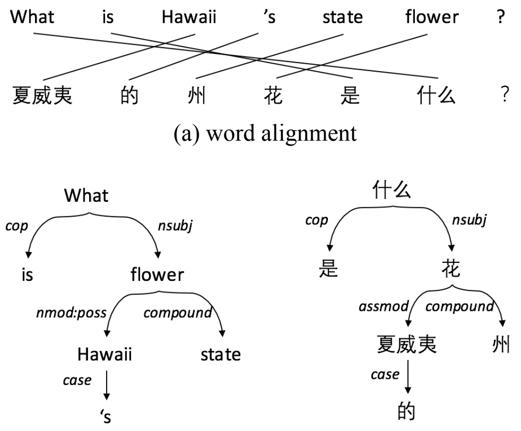
We use a bilingual dictionary to find bilingual word pairs. Each word w in the corpus has $1/2$ chance to be replaced by its counter-part word in the other language. If there are multiple translations for w in the bilingual dictionary, we randomly choose its replacement with probability $1/k$, where k is the number of translations for w in the bilingual dictionary.

In the bilingual dictionary, many translations are phrase based, for example, "how many" and "多少". Intuitively phrases should be treated as a whole and translated to words or phrases in the other languages. Otherwise, "how many" will be translated word by word as "如何很多", which makes no sense in Chinese. Therefore, we propose a simple method to learn phrase based bilingual word embeddings. When creating the artificial mixed language corpus, if we need to substitute a word with its translated phrase, we connect all the words in the phrase with underscores so that they can be treated as one unit during word embedding learning. We also preprocess the data by identifying all the phrases and concatenating all the words in the phrases that appear in the bilingual dictionary. We thus can learn phrase based bilingual embeddings.

The original English and Chinese corpora are still useful for encoding pure monolingual information. Therefore, we mix them together with the artificial mixed language corpus to form the final corpus for word embedding learning. In the data, phrases are also identified and connected using the same strategy. We use the CBOW model (Mikolov et al., 2013) for the bilingual word embedding learning. CBOW follows the assumption that similar words are more likely to appear in similar context. It casts word embedding learning into a word prediction problem given the context of the word. Because the CBOW model ignores word order within the window of contextual words, it may fail to capture the grammar or word order difference between two languages. We set a relatively larger CBOW window size (20) so that the window can cover an average sentence length. This is expected to ignore the grammar difference within a sentence and allow the CBOW model to learn bilingual word embeddings based on sentence level word co-occurrence.

2.2 Dependency grammar based CNN

Using the learned bilingual word embeddings as input, we adopt CNN for sentence modeling. When doing convolution and max pooling, each window is treated as a unit, therefore, only local words' relations are captured. Due to different



(a) word alignment

(b) dependency tree

Figure 1: An example to show dependency grammar can yield unified grammar between languages (Chinese and English).

grammars, local words' relation may vary across different languages. For example in Figure 1a, the basic CNN model will create six windows with size of 3 for each sentence: {Padding, Padding, What}, {Padding, What, is}, {What, is, Hawaii}, {is, Hawaii, 's}, {Hawaii, 's, state}, {'s, state, flower} for English, and {Padding, Padding, 夏威夷}, {Padding, 夏威夷, 的}, {夏威夷, 的, 州}, {的, 州, 花}, {州, 花, 是}, {花, 是, 什么} for Chinese. We can see that four windows in each sentence (out of six windows in that sentence) have different word ordering from the corresponding window in the other language.

To make relations captured in a window more meaningful for CNN, we adopt dependency based grammar as an intermediate grammar. As shown in Figure 1b, a dependency based CNN creates windows {What, ROOT, ROOT}, {is, What, ROOT}, {Hawaii, flower, What}, {'s, Hawaii, flower}, {state, flower, What}, {flower, What, ROOT} for English, and {夏威夷, 花, 什么}, {的, 夏威夷, 花}, {州, 花, 什么}, {花, 什么, ROOT}, {是, 什么, ROOT}, {什么, ROOT, ROOT} for Chinese. These dependency based windows capture similar word order and co-occurrence across languages. The order of the windows is not important as the max pooling layer ignores the global window order.

We therefore propose to incorporate dependency information into CNN. We evaluate the following three different setups.

(a) Dependency based CNN (DCNN): We adopt the dependency based CNN proposed by Ma et al. (2015), where instead of the natural word orders within a window, dependency based orders are used. For example, let x be the word embedding of current word w, then a dependency based window with size of 3 is $x \oplus Parent^1(x) \oplus Parent^2(x)$, where $Parent^1(x)$ and $Parent^2(x)$ are the embeddings of the parent and the grandparent of x respectively; $\oplus$ is concatenation operation. The dependency based windows will be passed through the convolution layer and max pooling layer and finally a softmax layer for classification. We use a window size of 3 (a short dependency path) here in order to make the model more robust across different languages.

(b) DCNN incorporating lexical features: Although dependency grammar is a good intermediate grammar, dependencies across languages are still not exactly the same. Second, dependency parsing is not perfect, especially for resource-poor languages. Therefore, it is possible that some word co-occurrence patterns cannot be captured. We thus add lexical features by adding an additional channel with window size equal to one, that is, each window has only one word. This lexicon input (a single word embedding) also passes through an independent convolution and pooling layer and the resulting feature is concatenated with the other abstract features.

(c) DCNN with phrase based grammar: In order to utilize phrase based bilingual embeddings, we make a modification in the dependency based CNN. If the input sentence contains a phrase in the bilingual dictionary, we combine the word nodes from the same phrase into a phrase node in the dependency tree. The combined phrase node will inherit all the parents and children from its contained word nodes. Then the phrase node will be treated as a single unit in the model.

3 Tasks and datasets

To evaluate our model, we select four sentence classification tasks including question classification, sentiment classification on movie review, sentiment classification on product review and sentence level event detection. For each task, we either use existing data or collect our own. It is difficult to find cross-lingual data with identical annotation schema for all the tasks. We thus collect English and Chinese corpora from tasks with similar annotation schema and take the overlapping part. For all the tasks, we train our model on English data, and test on Chinese data. To tune our model, we split Chinese dataset into validation and test sets.

Question classification (QC) aims to determine the category of a given question sentence. For English, we use the TREC[1] dataset. For Chinese, we use a QA corpus from HIT-IRLab[2]. We kept the six overlapped question types for both English and Chinese corpora. The final corpus includes 4,313 English questions and 4,031 Chinese questions (859 for testing, 859 for validation and 2,313 for training[3]).

Sentiment classification on movie review (SC-M) aims to classify a piece of given movie review into positive or negative. For English, we use IMDB polarity movie reviews from (Pang and Lee, 2004) (5,331 positive and 5,331 negative). For Chinese, we use the short Chinese movie reviews from Douban[4]. Like IMDB, users from Douban leave their comments along with a score for the movie. We collected 250 one star reviews (lowest score), and 250 five star reviews (highest score). We randomly split the 500 reviews into 200 for validation and 300 for testing.

Sentiment classification on product review (SC-P) aims to classify a piece of given product review into positive or negative. We use corpora from (Wan, 2011). Their Chinese dataset contains mostly short reviews. However, their English Amazon product reviews are generally longer, containing several sentences. Although our model is designed to take a single sentence as input, CNN can actually handle any input length. We remove reviews that are longer than 100 words and treat the remaining review as a single sentence. For dependency parsing, we combine the root of each sentence and make it a global dependency tree. In the end, we got 3,134 English product reviews (1,707 positive, 1,427 negative); 1000 (549 positive, 451 negative) and 314 (163 positive, 151 negative) Chinese ones for validation and testing respectively.

Sentence level event detection (ED) aims to determine if a sentence contains an event. ACE 2005 corpus[5] is ideal for cross-lingual tasks, because it contains annotated data for different languages with the same definition of events. Sentence is the smallest unit that contains a set of complete event information, i.e., triggers and corresponding arguments. To build the sentence level corpus, we first split document into sentences. For each sentence, if an event occurs (event triggers and arguments exist), we label the sentence as positive. Otherwise, we label it as negative. In the end we have 11,090 English sentences (3,688 positive, 7,402 negative). From the Chinese data we randomly selected 500 Chinese sentences (157 positive, 343 negative) for test, and 500 (138 positive, 362 negative) for validation. The remaining 5,039 ones (1767 positive, 3772 negative) are kept as training set. Because this is a detection task, we report F-score for it.

4 Experiment

We compare our bilingual word embedding based strategy to MT-based approach on the above four cross-lingual sentence classification tasks. Besides, we also evaluate the effectiveness of incorporating dependency information into CNN for sentence modeling.

4.1 Experiment setup

For the traditional MT-based cross-lingual method, we use the state-of-the-art statistical MT system Moses[6]. Language model is trained on Chinese gigawords corpus[7] with SRILM[8]. The parallel corpora used are from LDC[9]. We first translate English data into Chinese, and then apply the model trained on the translated dataset to the Chinese test data. For sentence classification, we use both basic CNN and DCNN (Ma et al., 2015).

We use monolingual word embeddings learned on the Chinese gigaword corpus for CNN and DCNN in MT-based method. For bilingual word embedding learning, we use "One Billion Word Language Modeling Benchmark"[10] and Chinese gigaword as unlabeled corpora for English and Chinese respectively. The bilingual dictionary is obtained from CC-CREDIT[11].

For CNN model training, we use the stochastic

[1] http://cogcomp.cs.illinois.edu/Data/QA/QC/

[2] http://ir.hit.edu.cn

[3] For QC and ED, we kept some samples as training set for an in-domain supervised model (refer to Section 4.2).

[4] http://www.douban.com

[5] http://projects.ldc.upenn.edu/ace/

[6] http://www.statmt.org/moses/

[7] https://catalog.ldc.upenn.edu/LDC2003T09

[8] http://www.speech.sri.com/projects/srilm/

[9] LDC: 2005T10, 2007T23, 2008T06, 2008T08, 2008T18, 2009T02, 2009T06, 2010T03

[10] http://www.statmt.org/lm-benchmark

[11] http://www.mandarintools.com/cedict.html

	QC	SC-M	SC-P	ED
	Accuracy			F-1
MT-based Method				
CNN	**83.00**	76.62	83.40	84.82
DCNN	82.89	75.97	81.50	84.40
Our Method With Bilingual Embedding				
CNN	68.10	64.29	64.80	82.06
DCNN	72.53	73.38	65.10	82.53
+Lex	79.28	75.00	78.60	83.17
+Lex+Phrase	82.19	**79.22**	**83.60**	**85.01**

Table 1: Results of different systems. +Lex: lexical features are used; +Phrase: phrase-based bilingual word embeddings and grammar are used.

gradient descent (SGD) learning method. We apply random dropout (Hinton et al., 2012) on the last fully connected layer for regularization. We use ADADELTA (Zeiler, 2012) algorithm to automatically control the learning rate and progress. The batch size for SGD and feature maps are tuned on the validation set for each task and fixed across different configurations. We preprocess all our corpora with Stanford CoreNLP (Manning et al., 2014), including word segmentation, sentence segmentation and dependency parsing.

4.2 Results

Table 1 shows the results of different systems. When using the MT based methods, the basic *CNN* achieves better results than *DCNN*. One possible reason is that the translation system produces errors, which may affect the performance of dependency parsing. For our method using bilingual word embeddings, basic *CNN* encodes only lexicon mapping information, and is not good at capturing grammar patterns. Therefore, it is natural this system has the lowest result. *DCNN* performs better than *CNN*, because it is able to capture additional grammar patterns across two languages by incorporating dependency information. Adding lexical features (*DCNN+Lex*) further improves performance. Given the fact that dependency parser is not perfect and dependency grammar between languages is not exactly the same, the grammar patterns that *DCNN* learned are not always reliable. The lexical feature here acts as an additional evidence to make the model more robust. *DCNN+Lex+Phrase* yields the best performance. The bilingual lexicon dictionary we use contains 54,168 Chinese words, and 29,355 of them have phrase-based translations (54.19%). Therefore, phrase-based bilingual word embeddings can represent sentences more accurately, and thus yield better results.

Compared to the MT-based approach, our cross-lingual model achieves comparable and even better performance. The advantage of our method is that we only use s dependency parser and bilingual dictionary, instead of a much more complicated machine translation system, which requires expertise knowledge about different languages, human designed features and expensive parallel corpus. Our method can be easily applied to any language pairs whose dependency parsers exist.

We further compare our cross-lingual model with a monolingual model for question classification and event detection. We have labeled Chinese training data for both tasks. We train a DCNN model on Chinese training data and then test on Chinese test set. For question classification, the monolingual model has an accuracy of 93.02%, and for event detection, its F-score is 87.28%. The event detection corpus has a consistent definition across two languages. Therefore, our cross-lingual system achieves close performance as the monolingual one. However, for question classification, the English and Chinese labeled data are constructed by two different teams and their annotation schemes are not identical. Therefore, the monolingual model performs much better than our cross-lingual model. Domain adaptation between two data sets may improve the performance for the bilingual model, but it is not the focus of this paper.

5 Conclusion

In this paper, we propose an efficient way to model cross-lingual sentences with only a bilingual dictionary and dependency parser. We evaluated our method on Chinese and English data and showed comparable and even better results than the traditional MT-based method on several sentence classification tasks. In addition, our method does not rely on expertise knowledge, human designed features and annotated resources. Therefore, it is easy to apply it to any language pair as long as there exist dependency parsers and a bilingual dictionary.

Acknowledgments

The work is partially supported by DARPA Contract No. FA8750-13-2-0041 and AFOSR award No. FA9550-15-1-0346.

References

Massih Amini, Nicolas Usunier, and Cyril Goutte. 2009. Learning from multiple partially observed views-an application to multilingual text categorization. In *Advances in neural information processing systems*, pages 28–36.

Zheng Chen and Heng Ji. 2009. Can one language bootstrap the other: a case study on event extraction. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 66–74. Association for Computational Linguistics.

Kevin Duh, Akinori Fujino, and Masaaki Nagata. 2011. Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 429–433. Association for Computational Linguistics.

Yuhong Guo and Min Xiao. 2012. Cross language text classification via subspace co-regularized multiview learning. *arXiv preprint arXiv:1206.6481*.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Xiao Ling, Gui-Rong Xue, Wenyuan Dai, Yun Jiang, Qiang Yang, and Yong Yu. 2008. Can chinese web pages be classified with english data source? In *Proceedings of the 17th international conference on World Wide Web*, pages 969–978. ACM.

Mingbo Ma, Liang Huang, Bowen Zhou, and Bing Xiang. 2015. Dependency-based convolutional neural networks for sentence embedding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 174–179. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127. Association for Computational Linguistics.

Leonardo Rigutini, Marco Maggini, and Bing Liu. 2005. An EM based training algorithm for cross-language text categorization. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pages 529–535. IEEE.

Chang Wan, Rong Pan, and Jiefei Li. 2011. Bi-weighting domain adaptation for cross-language text classification. In *Proceedings of International Joint Conference on Artificial Intelligence*, volume 22, pages 1535–1540.

Xiaojun Wan. 2011. Bilingual co-training for sentiment classification of chinese product reviews. *Computational Linguistics*, 37(3):587–616.

Min Xiao and Yuhong Guo. 2013. Semi-supervised representation learning for cross-lingual text classification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1465–1475. Association for Computational Linguistics.

Matthew D Zeiler. 2012. Adadelta: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.